IEEE ICASSP 2021

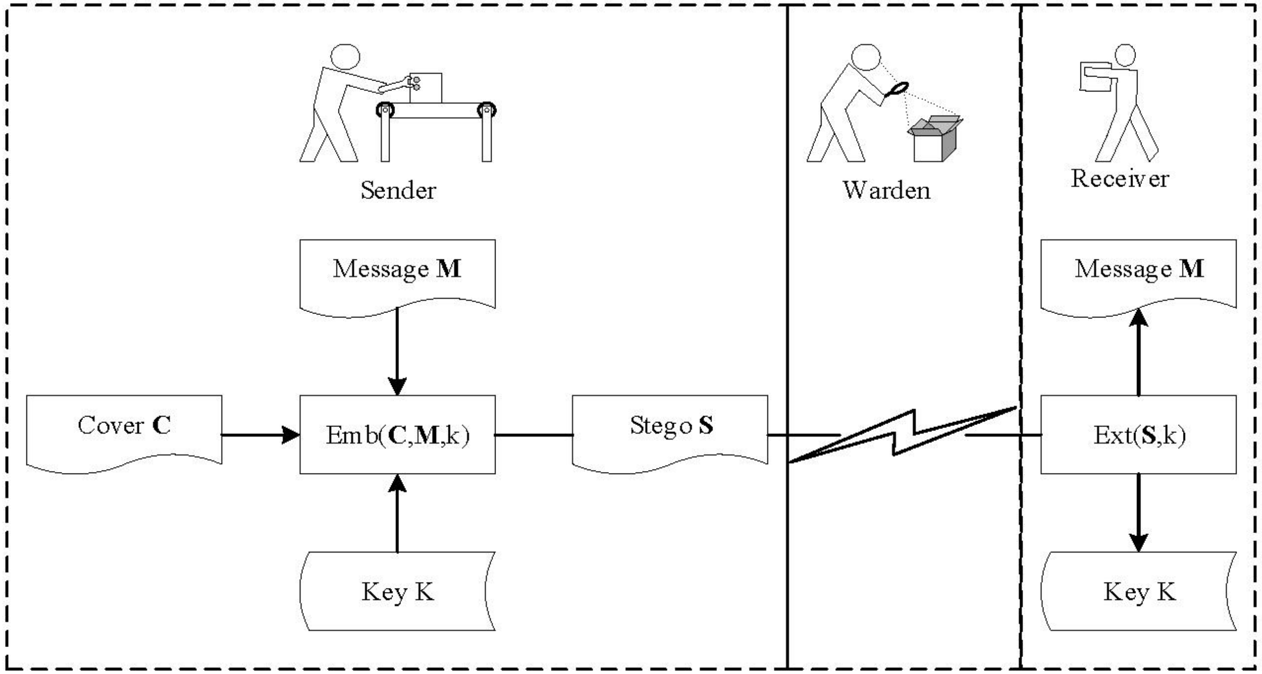
Image Steganography based on Iterative Adversarial perturbations onto A Synchronized-directions Sub-image

Xinghong Qin, Shunquan Tan, Weixuan Tang, Bin Li and Jiwu Huang

Shenzhen University

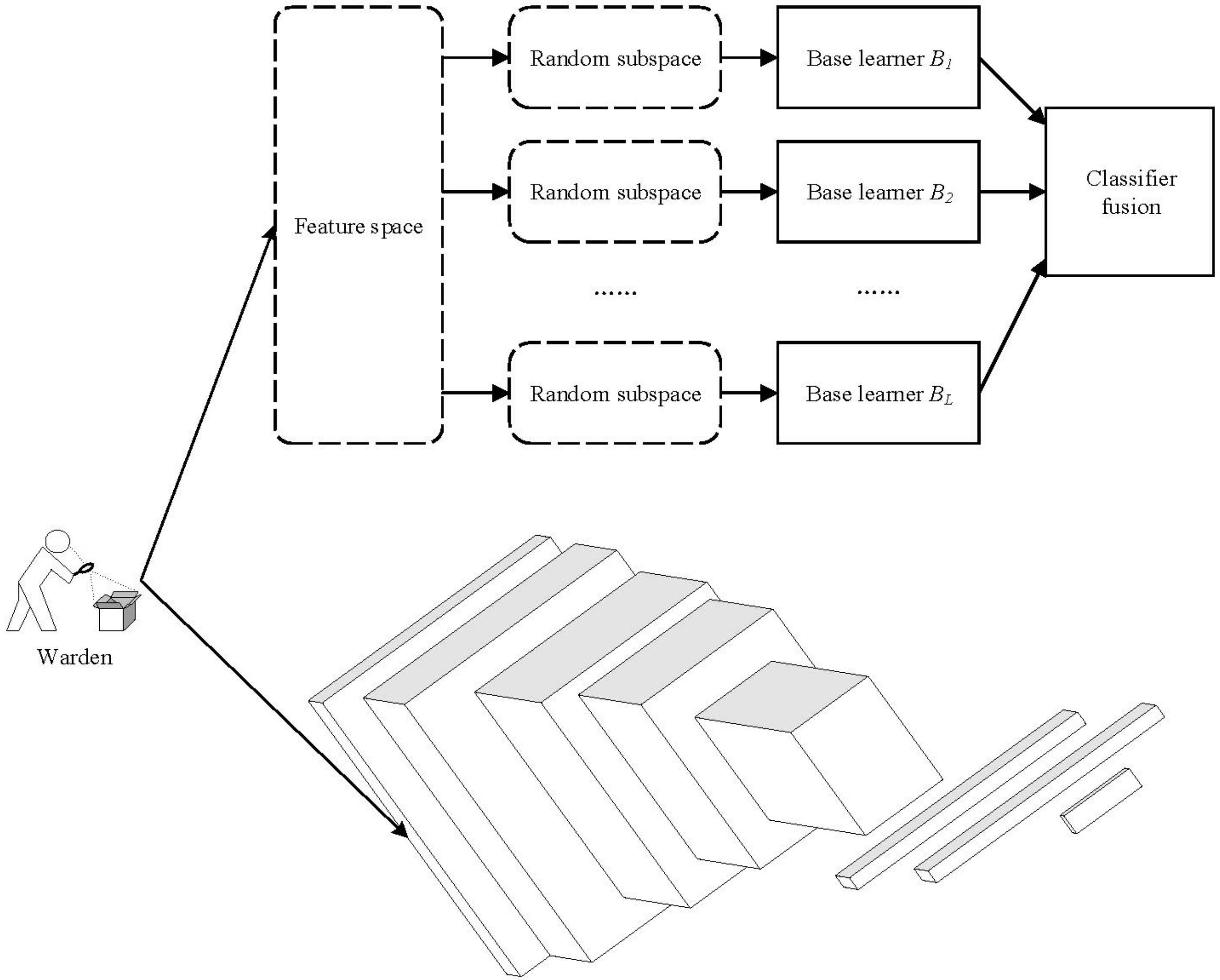
Introduction

- Steganography and steganalysis are a pair of antagonistic players.
  - Steganography:
    - Steganography is trying to escape being detected by steganalysis.
  - Steganalysis:
    - The warden discriminates whether a cover or a stego object is sent.
  - Scenario
    - The sender slightly modifies the cover **C** to conceal the secret message **M** to produce the stego **S**.
    - Send **S** to the receiver through the channel with passive the warden.
    - The receiver extracts **M** from the received **S**.
    - If the warden classifies the sent object is a stego, he maybe block-up the transmission or damage the sent object.



Introduction

- Steganography has to face challenges of both feature-based steganalysis and CNN steganalysis.



Introduction

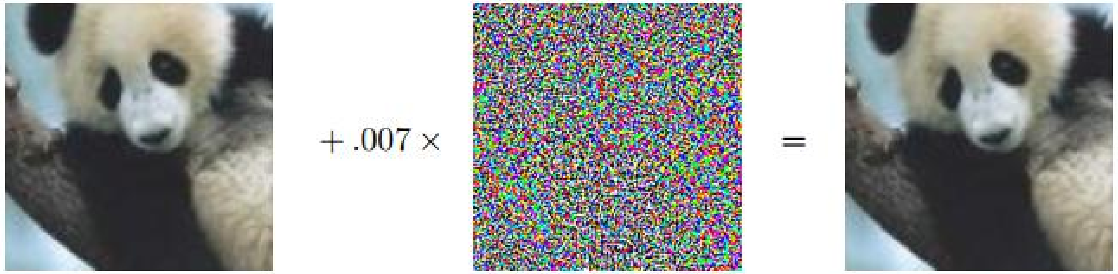
- Motivation.
  - Incorporate SMD strategy and adversarial examples to further enhance steganographic security to counter both feature-based steganalysis and CNN steganalysis.
    - Synchronizing modification directions (SMD) strategy can improve steganographic security.

$$D(\mathbf{X},\mathbf{Y}) = \sum_{(i,j),(k,l)\in\mathcal{C}} S_C(X_{ij}-Y_{i,j}, X_{kl}-Y_{kj}) \qquad (1)$$

$$S_C = \begin{array}{c|ccc} & -1 & 0 & 1 \\ \hline -1 & 0 & A_C & \nu A_C \\ 0 & A_C & 0 & A_C \\ 1 & \nu A_C & A_C & 0 \end{array}$$

  - Many machine learning classifiers are vulnerable to adversarial examples.

$$\mathbf{X}_{adv} = \mathbf{X} + \epsilon \cdot sign(\bigtriangledown_{\mathbf{X}} \mathscr{L}(\Phi(\mathbf{X}), \mathbf{y}_t)) \qquad (2)$$



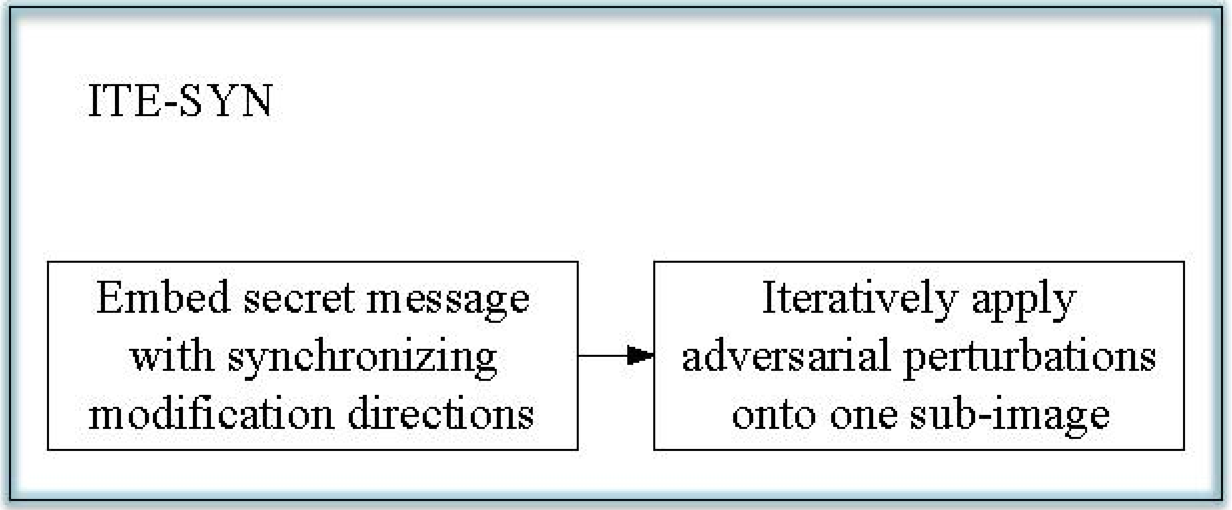
$\mathbf{X}$
"panda"
57.7% confidence

$sign(\bigtriangledown_{\mathbf{X}} \mathscr{L}(\Phi(\mathbf{X}), \mathbf{y}_t))$

$\mathbf{X}_{adv}$
"gibbon"
99.3% confidence

Our Method

- Base framework
  - ITE-SYN: ***ITE***ratively apply adversarial perturbations onto one ***SYN***chronized modification directions sub-image.



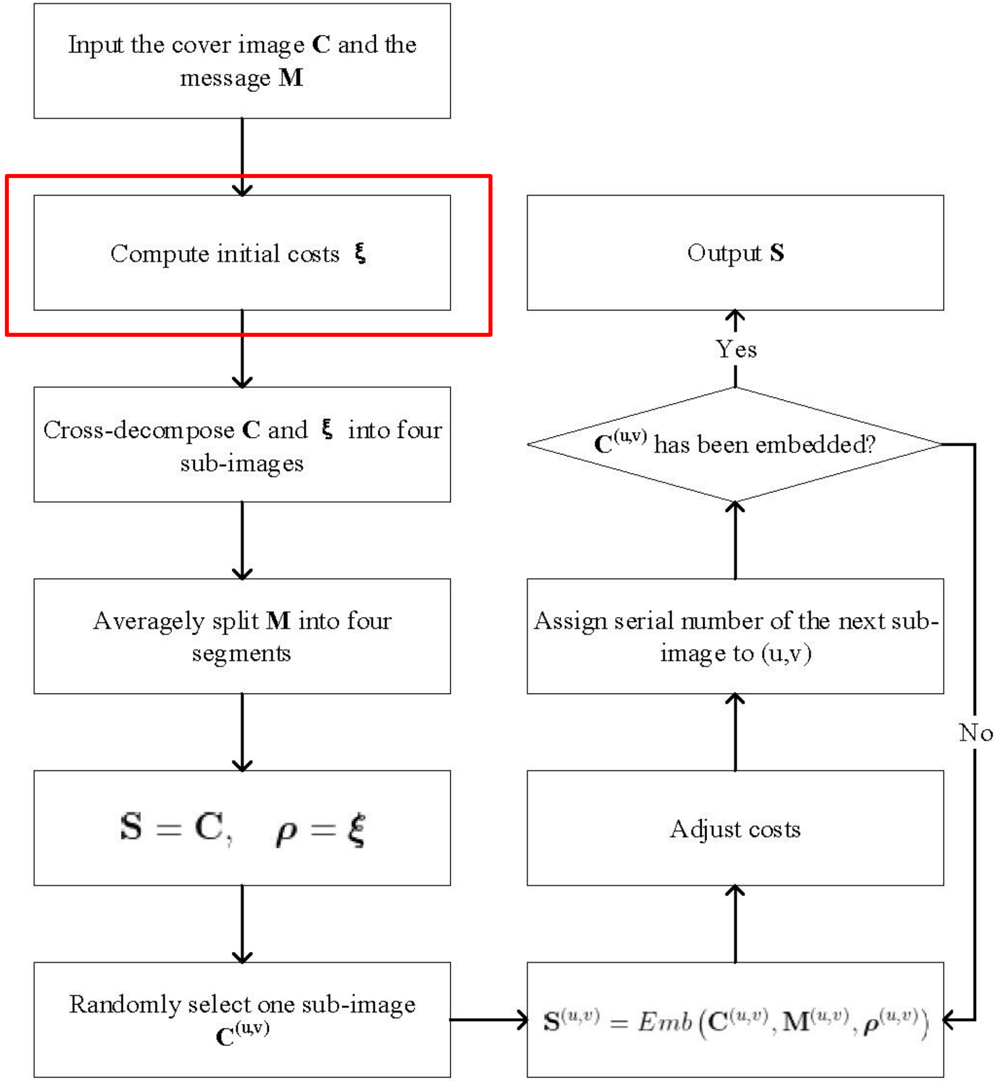
Our Method

- Embed secret message with synchronizing modification directions
  - Implement clustering modification directions (CMD) strategy.
    - The initial costs ξ are only computed once.
    - Adjust costs as

$$\begin{cases} \rho_{+}^{(i,j)} = \xi_{+}^{(i,j)}/\beta & \text{if } \sum_{\Delta c^{(r,s)} \in \mathcal{N}^{(i,j)}} \Delta c^{(r,s)} > 0, \\ \rho_{-}^{(i,j)} = \xi_{-}^{(i,j)}/\beta & \text{if } \sum_{\Delta c^{(r,s)} \in \mathcal{N}^{(i,j)}} \Delta c^{(r,s)} < 0, \\ \rho_{\pm}^{(i,j)} = \xi_{\pm}^{(i,j)} & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathcal{N}^{(i,j)} = \{\Delta c^{(r,s)} | r \in \{i-1, i+1\}, s \in \{j-1, j+1\}\}$
$\Delta \mathbf{C} = \mathbf{S} - \mathbf{C}$

- Select $\beta = 10$ for images with size-scale 256×256.



Our Method

- Iteratively apply adversarial perturbations.
  - We re-embed image to produce adversarial perturbations.

$$\begin{aligned}\Delta \mathbf{C}' = \mathbf{Z} - \mathbf{C} &= (\mathbf{Z} - \mathbf{S}) + (\mathbf{S} - \mathbf{C}) \\ &= \mathbf{n} + \Delta \mathbf{C},\end{aligned} \tag{4}$$

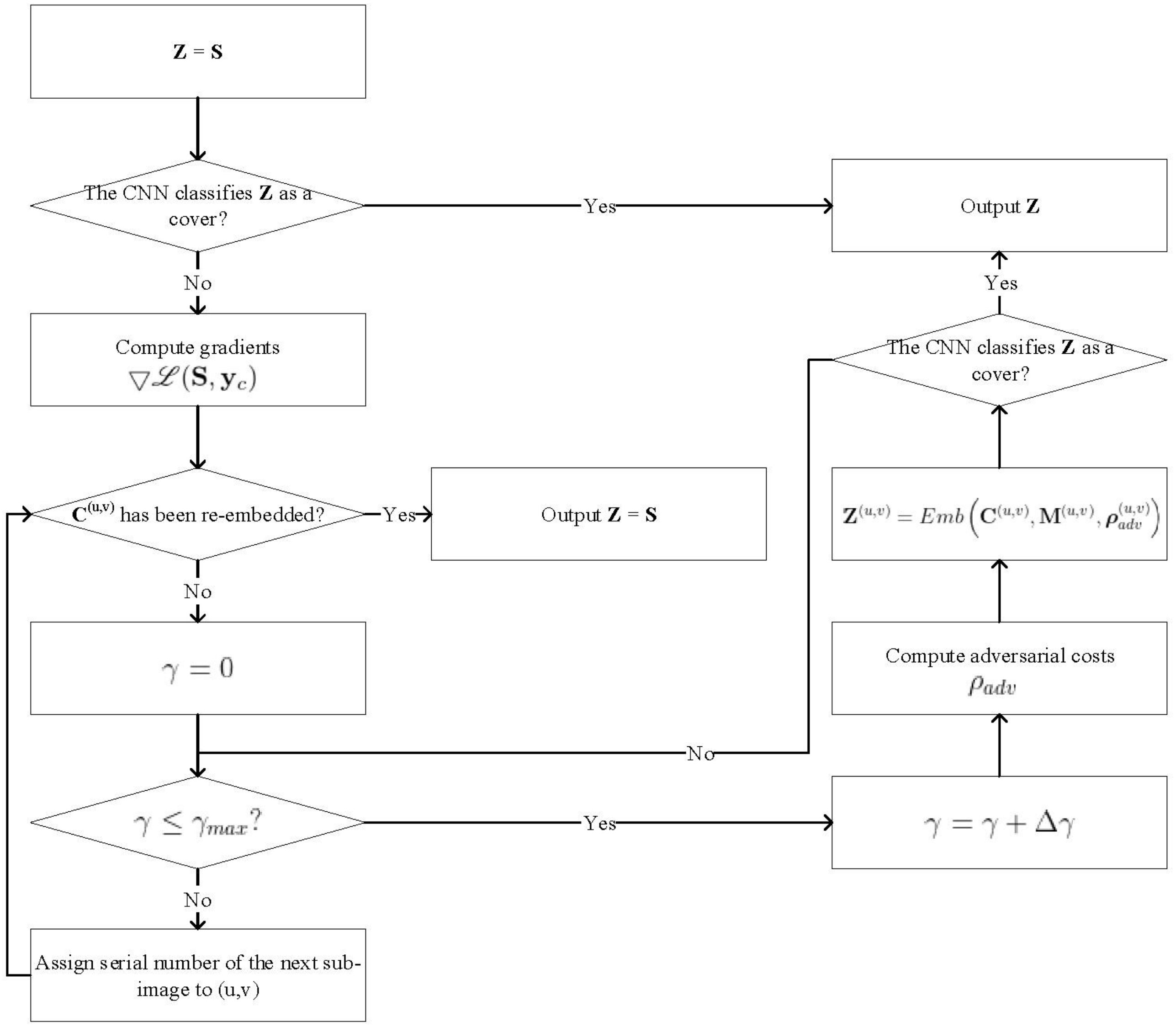
  - Adversarial costs are computed from embedding costs $\boldsymbol{\rho}$ adjusted by SMD.

$$\rho_{adv+}^{(i,j)} = \begin{cases} \rho_+^{(i,j)}(1+\gamma) & \text{if } \triangledown \mathscr{L}^{(i,j)}(\mathbf{S}, \mathbf{y}_c) > 0, \\ \rho_+^{(i,j)}/(1+\gamma) & \text{if } \triangledown \mathscr{L}^{(i,j)}(\mathbf{S}, \mathbf{y}_c) < 0, \\ \rho_+^{(i,j)} & \text{otherwise} \end{cases} \tag{5}$$

$$\rho_{adv-}^{(i,j)} = \begin{cases} \rho_-^{(i,j)}(1+\gamma) & \text{if } \triangledown \mathscr{L}^{(i,j)}(\mathbf{S}, \mathbf{y}_c) < 0, \\ \rho_-^{(i,j)}/(1+\gamma) & \text{if } \triangledown \mathscr{L}^{(i,j)}(\mathbf{S}, \mathbf{y}_c) > 0, \\ \rho_-^{(i,j)} & \text{otherwise} \end{cases} \tag{6}$$

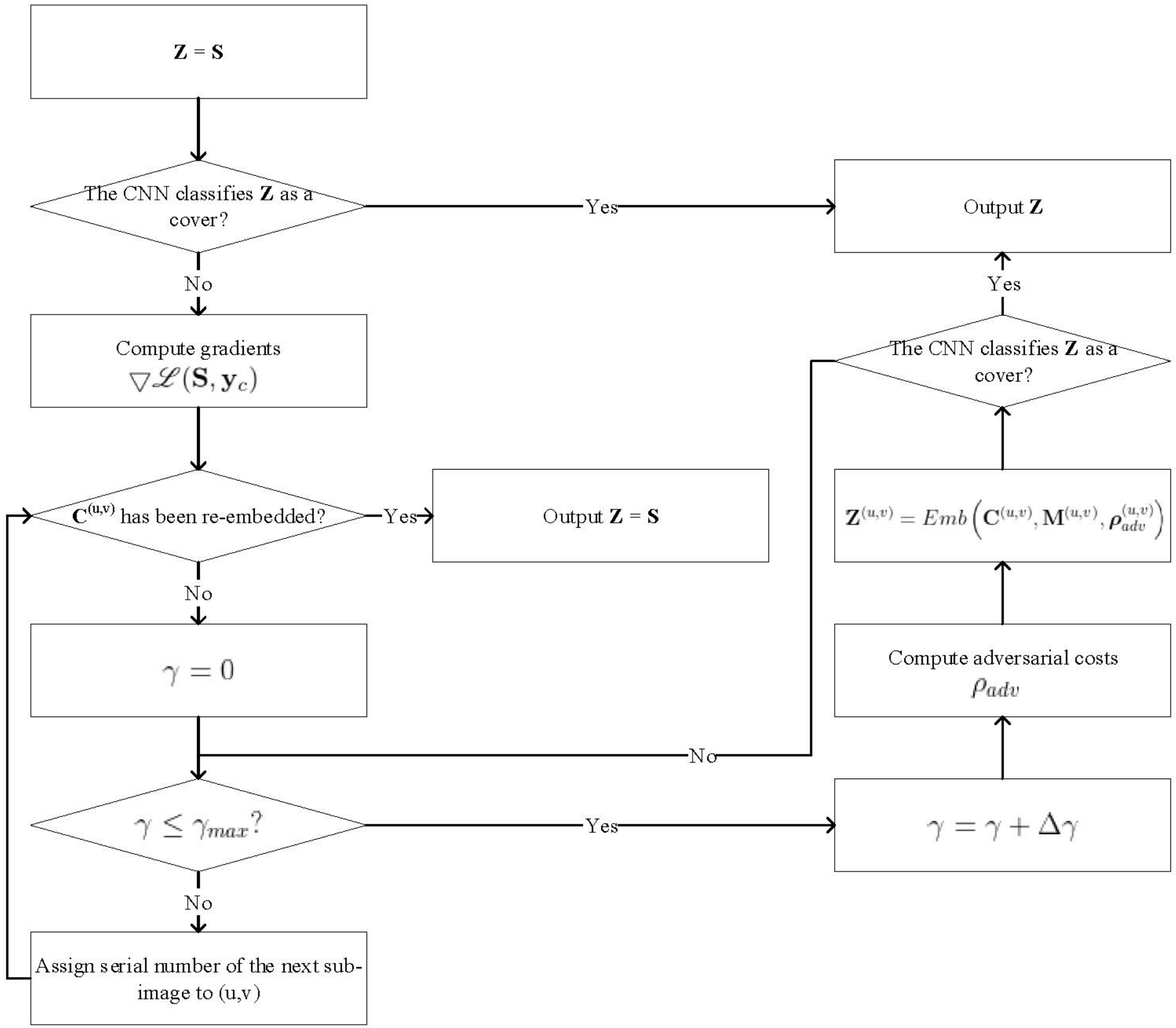
  - Parameters

$$\Delta\gamma = 0.1, \quad \gamma_{max} = 10.$$



Our Method

- Iteratively apply adversarial perturbations.
  - Adversarial perturbations are only applied onto one sub-image.
    - If re-embedding one sub-image is failed to deceive the target CNN classifier, the next sub-image will be selected to be re-embedded until all sub-images are tried re-embedding.



Experiments

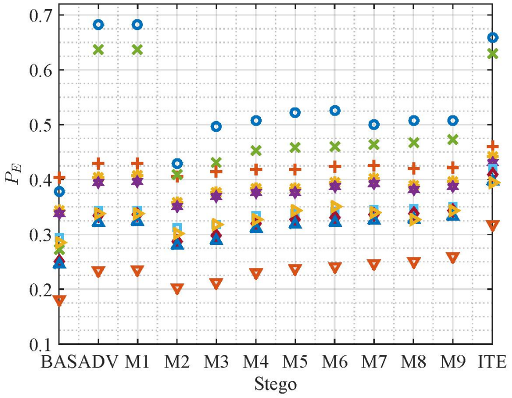
- Setup
  - Image database: BOSS256
    - Union of BOSSBase v1.01 and BOWS2. Totally 20000 images.
    - Resize each image from size-scale 512X512 to 256X256 by Matlab.
    - For CNN, 1000 images and 5000 images randomly selected from BOSSBase for validation and testing, other 14000 images are for training.
  - Cost functions
    - Heuristic method: HILL.
    - Model-based method: MiPOD.
  - Steganalysis
    - CNN classifiers
      - The target: XuNet, YeNet.
      - The non-target: SRNet.
    - Ensemble classifiers: SRM, maxSRMd2, PDASS.
  - Comparison schemes
    - ADV-EMB
    - MinMax + ADV-EMB.
  - Payload rates
    - 0.2 bpp and 0.4 bpp
  - Performance

$$P_E = \frac{P_{FA} + P_{MD}}{2} \tag{7}$$

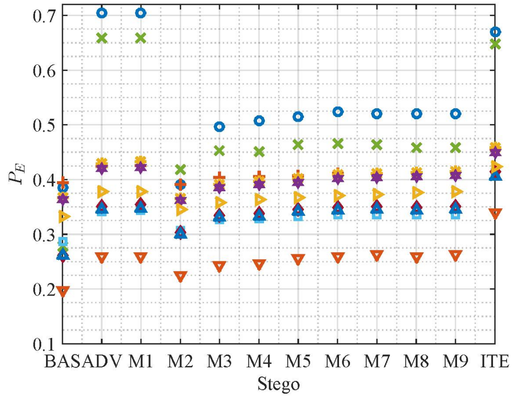
  - Stegos are created by the simulator unless specified.

Experiments

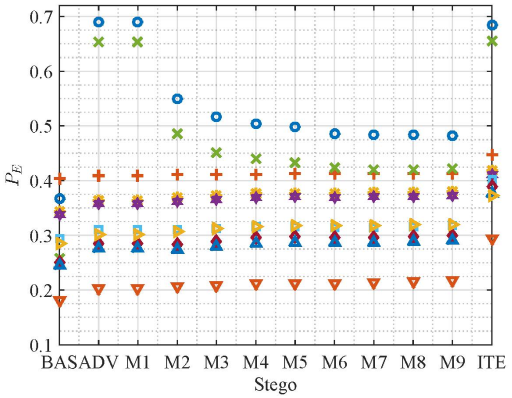
- Deceiving original classifiers
  - Notations
    - BAS: baseline.
    - ADV: ADV-EMB.
    - ITE: ITE-SYN.
    - M1-M9: versions of MinMax+ADV-EMB.
  - Target CNN classifier
    - XuNet: (a)-(b)
    - YeNet: (c)-(d)
  - Conclusion
    - ITE-SYN can effectively deceive the target CNN classifiers.
    - ITE-SYN improve steganographic performances to counter other original classifiers.



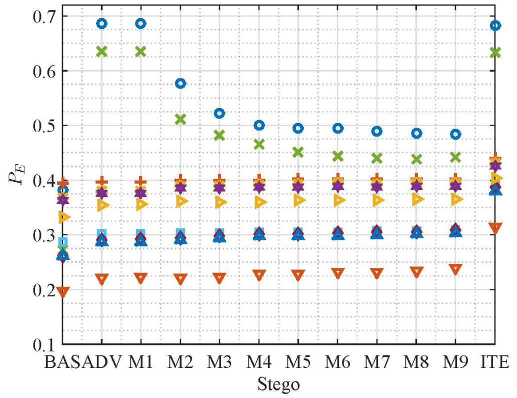
(a) HILL for XuNet



(b) MiPOD for XuNet



(c) HILL for YeNet



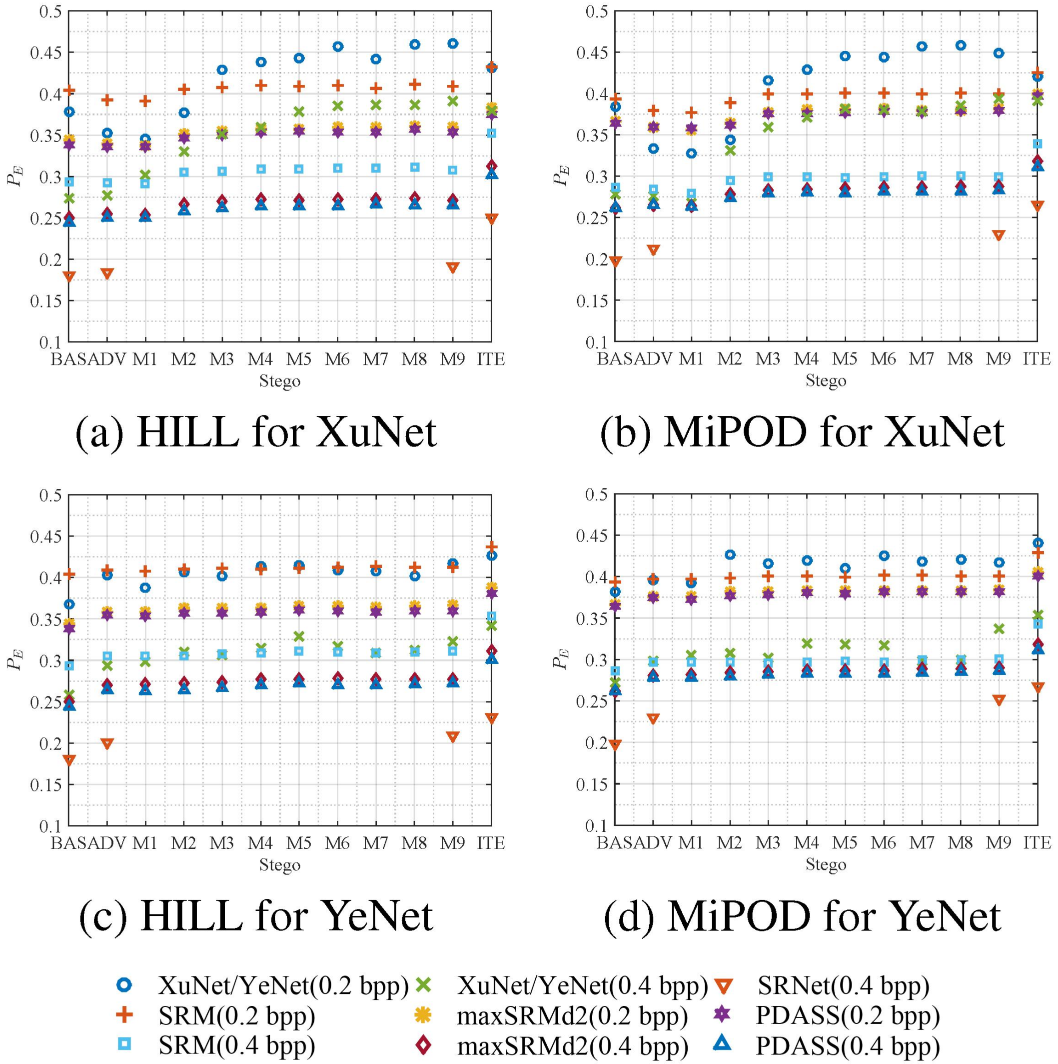
(d) MiPOD for YeNet

XuNet/YeNet(0.2 bpp) XuNet/YeNet(0.4 bpp) SRNet(0.2 bpp) SRNet(0.4 bpp)
SRM(0.2 bpp) maxSRMd2(0.2 bpp) PDASS(0.2 bpp)
SRM(0.4 bpp) maxSRMd2(0.4 bpp) PDASS(0.4 bpp)

Experiments

- Countering adversarial training classifiers
  - ITE-SYN outperforms ADV-EMB.
  - For comparison with MinMax+ADV-EMB,
    - ITE-SYN performs superior for non-target CNN classifiers and feature-based classifiers.
    - ITE-SYN performs superior when countering YeNet classifiers.
    - MinMax+ADV-EMB outperforms ITE-SYN after the fourth round when countering XuNet classifier.

- Discussion
  - Computational complexity of ITE-SYN is lower than of MinMax+ADV-EMB.
    - ITE-SYN creates only one stego image for each cover image.
  - It is predicted that steganographic performances of MinMax+ITE-SYN should be further improved.

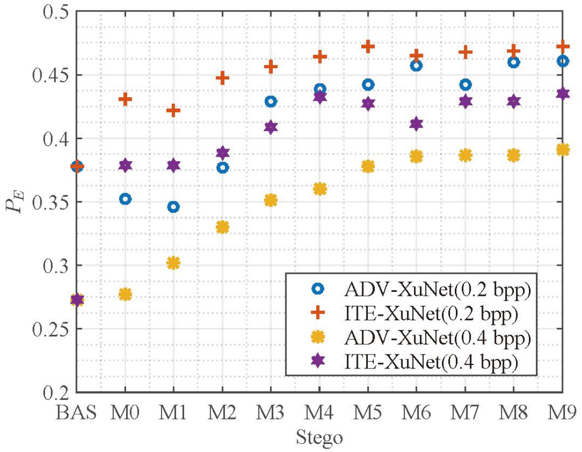


(a) HILL for XuNet (b) MiPOD for XuNet

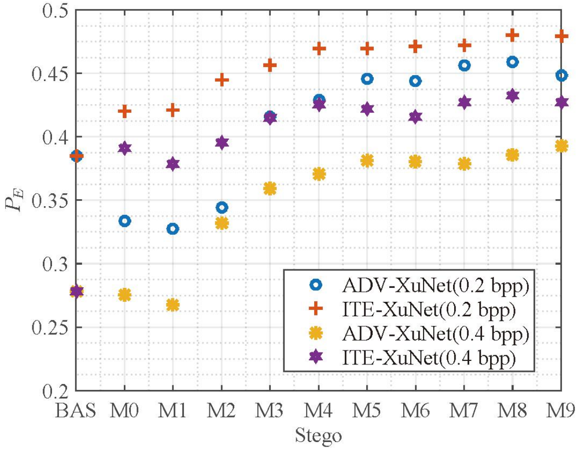
(c) HILL for YeNet (d) MiPOD for YeNet

Appendix: Issues

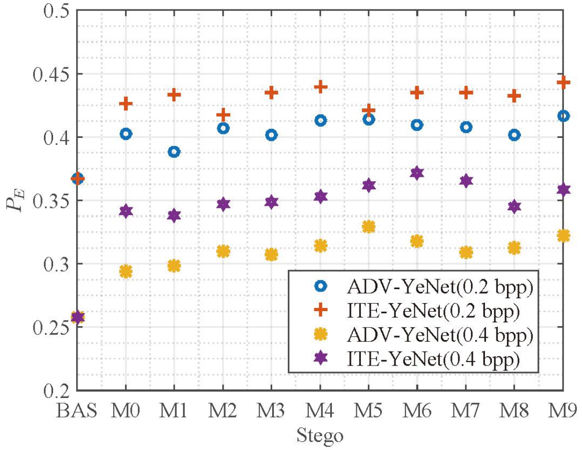
- Performances of MinMax+ITE-SYN
  - Notations
    - BAS: baseline.
    - M0-M9: rounds of MinMax.
  - Conclusion
    - MinMax+ITE-SYN outperforms MinMax+ADV-EMB.



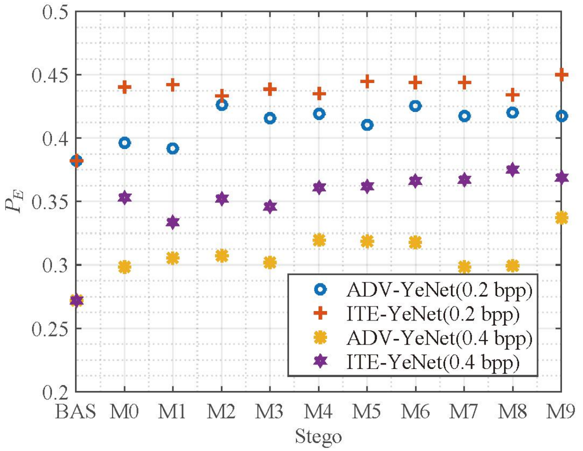
(a) HILL-XuNet



(b) MiPOD-XuNet



(c) HILL-YeNet



(d) MiPOD-YeNet

Performances of countering adversarial training classifiers.

Experiments

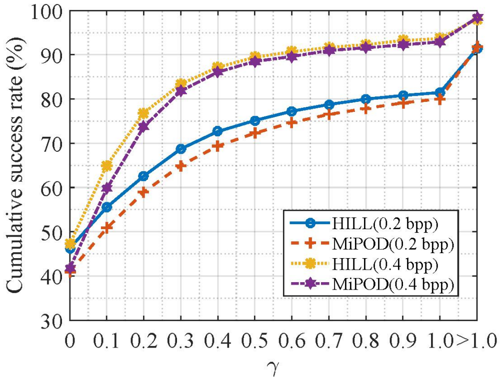
- Computational time (STCs)
  - Success rates are over 90%.
    - ITE-SYN can effectively deceive the target CNN classifiers.
  - Maximal iteration
    - ADV-EMB: 10.
    - ITE-SYN: 400.
  - Average computational times of ITE-SYN are less than of ADV-EMB, except for ITE-SYN for XuNet with payload rate 0.2 bpp.
    - Success rate of ITE-SYN is less about 5%.
- Cumulative success rate

$$\mathscr{P}(x_0) = P_r\{x \le x_0\} = \int_{-\infty}^{x_0} f(t)dt \qquad (8)$$

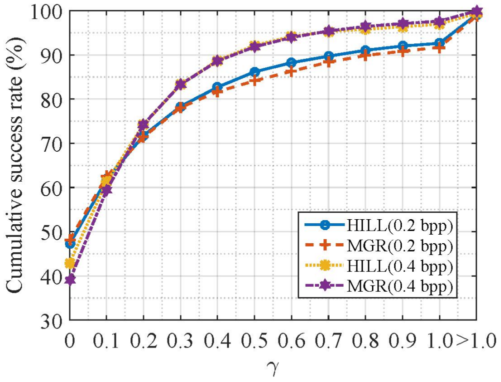
  - When $\gamma_{max} = 1$,
    - cumulative success rates are over 80%,
    - the maximal iteration of ITE-SYN: 40,
    - average time of creating adversarial stego image by ITE-SYN for XuNet as the target CNN classifiers with payload rate 0.2 bpp is 7.38 seconds.

Average success rate (in %) and computational time (in seconds) of creating an adversarial stego image.

Target	Scheme	0.2 bpp		0.4 bpp	
		Success rate	Time	Success rate	Time
XuNet	ADV-EMB	95.76	**10.40**	99.49	7.79
	ITE-SYN	90.79	24.19	97.99	**7.48**
YeNet	ADV-EMB	99.82	6.68	99.61	6.79
	ITE-SYN	98.95	**6.66**	99.60	**3.94**



(a) XuNet



(b) YeNet

Cumulative success rate of creating an adversarial stego images.

- Conclusion
  - Computational complexity of ITE-SYN is lower.

Conclusion

- ITE-SYN further enhances steganographic security countering both feature-based steganalysis and CNN steganalysis.
  - ITE-SYN can effectively deceive the target CNN classifiers, and can effectively resist on detection of other original classifiers, including both feature-base classifiers and CNN classifiers.
  - ITE-SYN has significant undetectability to counter adversarial training classifiers, including both feature-based classifiers and CNN classifiers.
  - Gradually increased adversarial perturbations are only applied onto one clustering modification directions sub-image.
    - It spends low computational expense.
    - It guarantees that adversarial perturbations applied are minimal.
    - It is unnecessary to search the optimal adversarial intensity.
- Future works
  - Extend the method to JPEG images.
    - Investigate incorporation of adversarial perturbations and effective cost strategy.
  - Investigate inner mechanisms of both SMD strategy and adversarial perturbations to design more powerful steganographic algorithm.

Thanks!